

DISTRIBUTED SOFT THRESHOLDING FOR SPARSE SIGNAL RECOVERY

C. Ravazzi, S. M. Fosson, E. Magli

Department of Electronics and Telecommunications, Politecnico di Torino, Italy

ABSTRACT

In this paper, we address the problem of distributed sparse recovery of signals acquired via compressed measurements in a sensor network. We propose a new class of distributed algorithms to solve Lasso regression problems, when the communication to a gateway node is not possible, e.g., due to communication cost or privacy reasons. More precisely, we introduce a distributed iterative soft thresholding algorithm (DISTA) that consists of three steps: an averaging step, a subgradient step, and a soft thresholding operation. We prove the convergence of DISTA in a network represented by a complete graph, and we show that it outperforms existing algorithms in terms of performance and complexity.

Index Terms— Distributed compressed sensing, distributed optimization, consensus algorithms, subgradient algorithms.

1. INTRODUCTION

Distributed compressed sensing [1] has recently emerged as a new research area that aims at decentralizing data acquisition and processing in compressed sensing. The rationale is the following: if we consider a network of sensors that individually acquire compressed measurements of correlated signals, we can expect a reduction in the number of measurements needed to obtain exact recovery. In fact, even if each sensor individually takes an insufficient number of measurements, reconstruction can be achieved by directing the whole network information to a single collection point, which includes a decoder that can reconstruct the signals in a joint fashion. However, in large-scale networks, gathering all data at a single point can be prohibitive from the energy consumption point of view, and can also introduce delays, severely reducing the sensor network performance. In other applications, agents providing private data may not be willing to share them [2].

The authors are with the Department of Electronics and Telecommunications (DET), Politecnico di Torino, Corso Duca degli Abruzzi 24, Torino, Italy. E-mails: sophie.fosson@polito.it, enrico.magli@polito.it, chiara.ravazzi@polito.it. The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement n 279848.

In this work, we consider the problem of in-network processing of information that is not centrally available at a gateway node, as formulated in [3]. Specifically, we assume that we are provided with a sensor network, in which sensors can store a limited amount of information, perform a low number of operations, and communicate under some constraints. Our aim is to study how to perform compressed acquisition and recovery leveraging on these seemingly scarce resources, with no computational support from an external gateway node. As is the case of decentralized methods, the key point is to suitably exploit local communication among sensors and develop an iterative algorithm that spreads the information through the network.

In particular, we propose a decentralized version of iterative thresholding methods [4], which basically consist of a subgradient step that seeks to minimize the Lasso functional, and a thresholding step that promotes sparsity. This is obtained by keeping the subgradient and thresholding steps, and adding a consensus step to share information among neighboring nodes. The reader can refer to [5–9] for an overview on consensus optimization problems.

As will be seen, the proposed distributed iterative soft thresholding algorithm (DISTA) not only does not require a centralized decoder, but also allows to dramatically reduce the number of measurements per sensor. In this paper we theoretically prove its convergence in fully connected networks and numerically verify its good performance, comparing it with that of existing methods, such as simultaneous orthogonal matching pursuit (SOMP) [1] and alternating direction method of multipliers (ADMM) [3, 10].

2. PROBLEM FORMULATION

2.1. Notation

Throughout this paper, we use the following notation. We denote column vectors with small letters, and matrices with capital letters. Given a matrix X , X^T denotes its transpose and $(X)_v$ (or x_v) denotes the v -th column of X . We consider \mathbb{R}^n as a Banach space endowed with the following norms:

$$\|x\|_p = \left(\sum_{i=1}^n |x_i|^p \right)^{1/p} \quad p = 1, 2.$$

For a rectangular matrix $X \in \mathbb{R}^{m \times n}$, we consider the Frobenius norm, which is defined as follows:

$$\|X\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n X_{ij}^2} = \sqrt{\sum_{j=1}^n \|(X)_j\|^2}.$$

A symmetric graph is a pair $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where \mathcal{V} is the set of vertices, and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of edges with the property that $(i, i) \in \mathcal{E}$ for all $i \in \mathcal{V}$ and $(i, j) \in \mathcal{E}$ implies $(j, i) \in \mathcal{E}$.

A matrix with non-negative elements P is said to be stochastic if $\sum_{j \in \mathcal{V}} P_{ij} = 1$ for every $i \in \mathcal{V}$. Equivalently, P is stochastic if $P\mathbf{1} = \mathbf{1}$. The matrix P is said to be adapted to a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ if $P_{v,w} = 0$ for all $(w, v) \notin \mathcal{E}$.

2.2. Model and assumptions

We consider a sensor network, whose topology is represented by a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. We assume that each node $v \in \mathcal{V}$ acquires linear measurements of the form

$$y_v = A_v x_0 + \xi_v \quad (1)$$

where $x_0 \in \mathbb{R}^n$ is k -sparse a signal (i.e., the number of its nonzero components is not larger than k), $\xi_v \in \mathbb{R}^m$ is an additive noise, and $A_v \in \mathbb{R}^{m \times n}$ (with $n \gg m$) is a random projection operator. If the measurements taken by all sensors were available at once in a single collection point that performs joint decoding, a solution to this problem would be to solve the basis pursuit denoising or Lasso problem [11, 12]. The Lasso refers to the minimization of the convex function $\mathcal{J} : \mathbb{R}^n \rightarrow \mathbb{R}$ defined by

$$\mathcal{J}(x, \lambda) := \sum_{v \in \mathcal{V}} \|y_v - A_v x\|_2^2 + \frac{2\lambda}{\tau} \|x\|_1 \quad (2)$$

where $\lambda > 0$ is a scalar regularization parameter that is usually chosen by cross validation [13] and $\tau > 0$. Let us denote the solution of (2) as

$$\hat{x} = \hat{x}(\lambda) = \operatorname{argmin}_{x \in \mathbb{R}^N} \mathcal{J}(x; \lambda). \quad (3)$$

This optimization problem is shown to provide an approximation with a bounded error, which is controlled by α [14].

A large amount of literature has been devoted to developing fast algorithms for solving the optimization problem (2) and characterizing the performance and optimality conditions. We refer to [4] for an overview of these methods.

2.3. Iterative soft thresholding

A popular approach to solve the optimization problem in (2) is the iterative soft thresholding algorithm (ISTA). ISTA is based on moving at each iteration in the direction of the steepest descent and thresholding to promote sparsity [15].

Let us collect the measurements in the vector $y = (y_1^T, \dots, y_{|\mathcal{V}|}^T)^T$ and let A be the complete sensing matrix $A = (A_1^T, \dots, A_{|\mathcal{V}|}^T)^T$. Given $x(0)$, iterate for $t \in \mathbb{N}$

$$x(t+1) = \eta_\lambda(x(t) + \tau A^T(y - Ax))$$

where τ is the stepsize in the direction of the steepest descent. The operator η is a thresholding function to be applied elementwise, i.e. $\eta_\lambda(x) = \operatorname{sgn}(x)(|x| - \lambda)$ if $|x| < \lambda$ and $\eta_\lambda(x) = 0$ otherwise.

The convergence of this algorithm was proved in [15], under the assumption that $\|A\|^2 < 1/\tau$.

3. PROPOSED DISTRIBUTED ALGORITHM

As has been said, transmitting all data collected in a sensor network to a centralized unit for joint decoding is not an efficient approach. In this work we propose a distributed iterative algorithm to approximate x_0 , in which the agents only exchange information with their nearest neighbors at each iteration, without any central coordination.

In particular, we describe a family of simple, easy to implement, relaxed subgradient thresholding methods, and prove their convergence.

3.1. A consensus-based reformulation of the Lasso

We recast the optimization problem in (2) into a separable form which facilitates distributed implementation. The goal is to split this problem into simpler subtasks executed locally at each node.

Let us replace the global variable x in (2) with local variables $\{x_v\}_{v \in \mathcal{V}}$, representing estimates of x_0 , provided by each node. While the conventional centralized Lasso problem attempts to minimize $\mathcal{J}(x, \alpha)$, we recast the distributed problem as an iterated minimization of the functional $\mathcal{F} : \mathbb{R}^{n \times |\mathcal{V}|} \mapsto \mathbb{R}^+$ defined as follows

$$\mathcal{F}(x_1, \dots, x_{|\mathcal{V}|}) = \sum_{v \in \mathcal{V}} \left[\|y_v - A_v x_v\|_2^2 + \frac{2\lambda}{\tau |\mathcal{V}|} \|x_v\|_1 + \frac{1-\gamma}{2\tau\gamma} \sum_{w \in \mathcal{V}} P_{v,w} \|x_w - x_v\|^2 \right] \quad (4)$$

where $P = [P_{v,w}]_{v,w \in \mathcal{V}}$ is a stochastic matrix adapted to the graph \mathcal{G} , and $\gamma \in (0, 1)$ is a parameter.

By minimizing \mathcal{F} , each node seeks to recover the sparse vector x_0 from its own linear measurements, and to enforce agreement with the estimates calculated by other sensors in the network. It should also be noted that if there is consensus, in the sense $x_v = \bar{x}$ for all $v \in \mathcal{V}$, then $\mathcal{F}(\bar{x}, \dots, \bar{x}) = \mathcal{J}(\bar{x}, \alpha)$.

Note that γ can be viewed as a temperature parameter; as γ decreases, estimates x_v associated with adjacent nodes become increasingly correlated. Let us denote $\{\hat{x}_v^\gamma\}_{v \in \mathcal{V}}$ the solution of (4). If \mathcal{G} is connected, then we expect that $\lim_{\gamma \rightarrow 0} \hat{x}_v^\gamma = \hat{x}, \forall v \in \mathcal{V}$.

3.2. DISTA: algorithm description

The information state of a node v at time t , denoted as $x_v(t)$, is an estimate of an optimal solution to the problem (2). The update at time $t + 1$ is obtained by combining this current estimate with the ones received from some of the other agents.

Given a strongly connected symmetric graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, let P be the Metropolis random walk transition matrix (see [16]): if $i \neq j$

$$P_{ij} = \begin{cases} 0 & \text{if } (i, j) \notin \mathcal{E} \\ (\max\{\deg(i) + 1, \deg(j) + 1\})^{-1} & \text{if } (i, j) \in \mathcal{E} \end{cases}$$

where $\deg(i)$ denotes the degree (the number of neighbors) of unit i in the graph \mathcal{G} ; and $P_{ii} = 1 - \sum_{j \neq i} P_{ij}$. Each node v has a message stored in its memory at time t , denoted as $x_v(t)$.

DISTA: Given $x_v(0) = 0$, iterate for $t \in \mathbb{N}$

$$\begin{aligned} \bar{x}_v(t) &= \sum_{w \in \mathcal{V}} P_{v,w} x_w(t), \\ u_v(t) &= x_v(t) + \tau A_v^T (y_v - A_v x_v(t)), \\ x_v(t+1) &= \eta_\alpha ((1 - \gamma) \bar{x}_v(t) + \gamma u_v(t)). \end{aligned}$$

with $\gamma \in (0, 1)$ and $\alpha = \lambda/(|\mathcal{V}|\gamma)$.

It should be noted that if $|\mathcal{V}| = 1$, DISTA coincides with ISTA.

3.3. Discussion and comparison with related work

A few contributions towards distributed reconstruction are available in the literature.

In particular, simultaneous orthogonal matching pursuit (SOMP) assumes that each sensor measures a different signal, but all signals have a common sparse support. The algorithm first estimates the support by averaging the information held by the nodes, then runs an individual recovery procedure at each node. We notice that the averaging step could be easily performed in a distributed way on networks with communication constraints, using classical consensus methods [5]. On the other hand, in SOMP, after the averaging step, the signal is recovered separately by the sensors, while in DISTA cooperation is exploited also for reconstruction. Even if the measured signal is assumed to be common, this cannot be practically used in SOMP, as the recovery procedure is not iterative.

Another algorithm for distributed sparse linear regression is the ADMM [3, 10], which tackles the problem in (4) by introducing dual variables and minimizing the augmented Lagrangian in an iterative way with respect to the primal and dual variables. The algorithm entails the following steps for each $t \in \mathbb{N}$: agent v receives the local estimates from its neighbors, uses them to evaluate the dual price vector and the new estimate via coordinate descent and thresholding operation. The

tricky point of this algorithm is the inversion of an $n \times n$ matrix at each node, which may be computationally demanding for very large n . Compared to ADMM, the updates in DISTA are extremely simple and involve just scaling and addition of vectors and soft thresholding operations.

4. MAIN CONTRIBUTION

4.1. Theoretical results

Let us define the operator $\Gamma : \mathbb{R}^{n \times |\mathcal{V}|} \mapsto \mathbb{R}^{n \times |\mathcal{V}|}$ where

$$(\Gamma X)_v := \eta_\alpha [(1 - \gamma) \bar{x}_v + \gamma(x_v(t) + \tau A_v^T (y_v - A_v x_v(t)))]$$

and $v \in \mathcal{V}$. DISTA can be rewritten as

$$X(t+1) = \Gamma X(t)$$

with any initial condition $X(0)$.

The following theorem ensures the convergence of DISTA.

Theorem 1. *If \mathcal{G} is complete and $\tau < \|A_v\|_2^{-2}$ for all $v \in \mathcal{V}$, the following hold for any initial choice $X(0)$:*

1. *there exists $X^* \in \mathbb{R}^{n \times |\mathcal{V}|}$ such that $\Gamma X^* = X^*$;*
2. *DISTA produces a sequence $\{X(t)\}_{t \in \mathbb{N}}$ such that*

$$\lim_{t \rightarrow \infty} \|X(t) - X^*\|_F = 0$$

3. *the limit point X^* is a minimizer of \mathcal{F} .*

Sketch of the proof. Following [15], the sequence of the $\{X(t)\}_{t \in \mathbb{N}}$ is proved to converge to a fixed point of Γ , by applying the Opial's Theorem [17]; then the equivalence of fixed points Γ and minimizers of $\mathcal{F}(X)$ is obtained by standard variational techniques. For brevity, the complete proof is deferred to [18]. \square

4.2. Numerical results

To demonstrate the performance of DISTA, we conduct a series of experiments for the complete graph architecture and for a variety of total number of measurements. We consider the complete topology where $P_{ij} = \frac{1}{N}$ for every $i, j = 1, \dots, N$. For a fixed n , we construct random recovery scenarios for sparse vector x_0 . For each n , we vary the number of measurements m per node and the number of nodes in the network. For each $(N, m, |\mathcal{V}|)$ triple, we repeat the following procedure 50 times.

A signal is generated by choosing k nonzero components uniformly among the n elements and sampling the entries from a Gaussian distribution $\mathcal{N}(0, 1)$. Matrices $(A_v)_{v \in \mathcal{V}}$ are sampled from the Gaussian ensemble with m rows, n columns, null mean and variance $\frac{1}{m}$. We fix $n = 150$, $k = 15$, $\alpha = 10^{-4}$, and $\tau = 0.02$.

In the noise-free case, we show the performance of DISTA in terms of reconstruction probability as a function of the

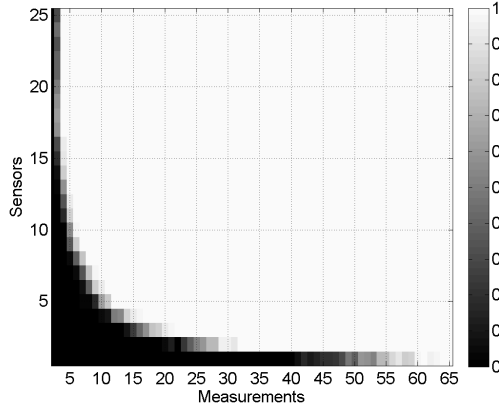


Fig. 1. Noise-free case: performance analysis of DISTA for complete graph, $n = 150$, $k = 15$.

number of measurements (see Figure 1). In particular, we declare x_0 to be recovered if $\sum_{v \in \mathcal{V}} \|x_0 - x_v^*\|_2^2 / (n|\mathcal{V}|) < 10^{-4}$. The color of the cell in the figures reflects the empirical recovery rate of the 50 runs (scaled between 0 and 1). White denotes perfect recovery in all experiments, and black denotes failure for all experiments. It should be noted that the number of total measurements $m|\mathcal{V}|$, which are sufficient for successfully recovery, remains constant.

In Figure 2 the probability of recovery of DISTA and SOMP (obtained with 50 runs) are compared as a function of the number of measurements per sensor. The curves are obtained for different number of sensors. SOMP is assumed to know the sparsity value $k = 15$. We immediately notice that the number of measurements needed for success by DISTA is smaller. Indeed, for SOMP, a number of measurements not smaller than k is a necessary (but not sufficient) condition for good recovery. This is evident in Figure 2, which shows that there are no recovery occurrences below $k = 15$, while above this threshold the probability of recovery increases with the dimension of the network. This is a substantial drawback that DISTA is able to overcome.

Finally, let us consider the noise case. In Figure 3, the mean square error

$$\text{MSE} = \frac{\sum_{v \in \mathcal{V}} \|x_0 - x_v^*\|_2^2}{n|\mathcal{V}|},$$

averaged over 50 runs, is plotted as a function of the signal-to-noise ratio

$$\text{SNR} = \frac{\mathbb{E} [\sum_{v \in \mathcal{V}} \|y_v\|^2]}{\mathbb{E} [\sum_{v \in \mathcal{V}} \|\zeta_v\|^2]}$$

for both DISTA and SOMP. The number of sensors is $|\mathcal{V}| = 10$. For equal SNR, the MSE decreases as the number m of measures for node increases. It should be noted that DISTA performs better than SOMP: $m = 6$ measures are sufficient for DISTA to obtain a MSE lower than the one obtained by SOMP with $m = 18$ measures.

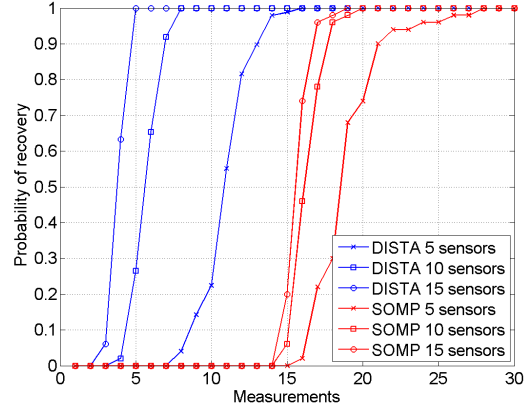


Fig. 2. Noise-free case: DISTA vs SOMP, complete graph, $n = 150$, $k = 15$.

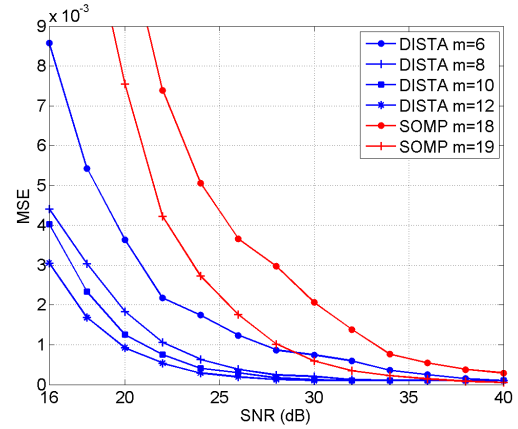


Fig. 3. Noise case: DISTA vs SOMP, complete graph, $n = 150$, $k = 15$, $|\mathcal{V}| = 10$.

In results not reported here for brevity, we have observed that the performance of DISTA is not strongly affected by the graph topology; this suggests that decentralization is not a drawback.

5. CONCLUDING REMARKS

The problem of distributively estimating sparse signals from compressed measurements in sensor networks with limited communication capability is studied. In particular, the DISTA algorithm has been proposed. The main contribution includes the proof of convergence of the algorithm to a local minimum of the distributed Lasso estimator. We also show simulation results showing that DISTA significantly outperforms SOMP.

6. REFERENCES

- [1] D. Baron, M. F. Duarte, M. B. Wakin, S. Sarvotham, and R. G. Baraniuk, "Distributed compressive sensing," <http://arxiv.org/abs/0901.3403>, 2005 (revised 2009).
- [2] P. A. Forero, A. Cano, and G. B. Giannakis, "Consensus-based distributed support vector machines," *J. Mach. Learn. Res.*, vol. 99, pp. 1663–1707, August 2010.
- [3] G. Mateos, J. A. Bazerque, and G. B. Giannakis, "Distributed sparse linear regression.," *IEEE Transactions on Signal Processing*, vol. 58, no. 10, pp. 5262–5276, 2010.
- [4] M. Fornasier, *Theoretical Foundations and Numerical Methods for Sparse Recovery*. Radon Series on Computational and Applied Mathematics, 2010.
- [5] J. N. Tsitsiklis, *Problems in Decentralized Decision Making and Computation*. PhD thesis, Department of EECS, MIT, November 1984.
- [6] R. Olfati-saber, J. A. Fax, and R. M. Murray, "Consensus and cooperation in networked multi-agent systems," in *Proceedings of the IEEE*, p. 2007, 2007.
- [7] F. Bullo, J. Cortes, and S. Martinez, *Distributed Control of Robotic Networks*. Applied Mathematics Series, 2009.
- [8] S. S. Ram, A. Nedic, and V. V. Veeravalli, "Distributed subgradient projection algorithm for convex optimization," in *ICASSP*, pp. 3653–3656, IEEE, 2009.
- [9] A. Nedic, A. Ozdaglar, and P. A. Parrillo, "Constrained consensus and optimization in multi-agent networks," *IEEE Transactions on Automatic Control*, vol. 55, pp. 922–938, 2010.
- [10] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, pp. 1–122, Jan. 2011.
- [11] S. S. Chen, D. L. Donoho, Michael, and A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, pp. 33–61, 1998.
- [12] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288, 1994.
- [13] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer Series in Statistics, New York, NY, USA: Springer New York Inc., 2001.
- [14] E. J. Candés, "The restricted isometry property and its implications for compressed sensing," *Compte Rendus de l'Academie des Sciences. Paris, France, ser. I*, vol. 346, pp. 589–592, 2008.
- [15] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Comm. Pure Appl. Math.*, vol. 57, no. 11, pp. 1413–1457, 2004.
- [16] J.-J. Xiao, A. Ribeiro, Z.-Q. Luo, and G. Giannakis, "Distributed compression-estimation using wireless sensor networks," *Signal Processing Magazine, IEEE*, vol. 23, pp. 27 – 41, july 2006.
- [17] Z. Opial, "Weak convergence of the sequence of successive approximations for nonexpansive mappings," *Bull. Amer. Math. Soc.*, vol. 73, pp. 591–597, 1967.
- [18] C. Ravazzi, S. M. Fosson, and E. Magli, "DISTA convergence – Complete graphs." <http://calvino.polito.it/~ravazzi/publications/Report.pdf>, 2012.